Learning Unbiased Representations via Mutual Information Backpropagation

Ruggero Ragonesi[1,2], Riccardo Volpi[3], Jacopo Cavazza[1], Vittorio Murino[1,4,5]

[1] PAVIS Department, Istituto Italiano di Tecnologia, Genova, Italy
[2] DITEN Department, University of Genova, Italy
[3] Naver Labs Europe, Grenoble, France
[4] Department of Computer Science, University of Verona, Italy
[5] Huawei Technologies Ltd., Ireland Research Center, Dublin, Ireland
name.lastname@iit.it

Abstract

We are interested in learning data-driven representations that can generalize well, even when trained on inherently biased data. In particular, we face the case where some attributes (bias) of the data, if learned by the model, can severely compromise its generalization properties. We tackle this problem through the lens of information theory, leveraging recent findings for a differentiable estimation of mutual information. We propose a novel end-to-end optimization strategy, which simultaneously estimates and minimizes the mutual information between the learned representation and specific data attributes. When applied on standard benchmarks, our model shows comparable or superior classification performance with respect to state-of-the-art approaches. Moreover, our method is general enough to be applicable to the problem of "algorithmic fairness", with competitive results.

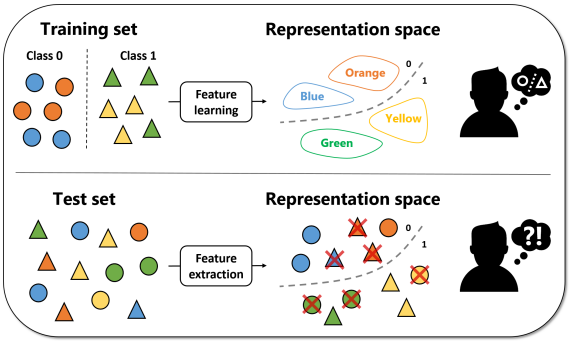

Figure 1: **Problem setting.** When learning a feature representation from the data itself *(top)*, we may undesirably capture the inherent bias of the dataset (here, exemplified by colors), as opposed to learning the desired patterns (here, represented by shapes). This results in models that poorly generalize when deployed into unbiased scenarios *(bottom)*.

1. Introduction

The need for proper data representations is ubiquitous in machine learning and computer vision [7]. Indeed, given a learning task, the competitiveness of the proposed models crucially depends upon the data representation one relies on. In the last decade, the mainstream strategy for designing feature representations switched from hand-crafting to learning them in a data-driven fashion [11, 25, 40, 32, 18, 19, 45]. In this context, deep neural networks have shown an extraordinary efficacy in learning hierarchical representations via backpropagation [37]. However, while learning representations from data allows achieving remarkable results in a broad plethora of tasks, it leads to the following shortcoming: a representation may inherit the intrinsic bias of the dataset used for training.

This is highly undesirable, because it leads a model to poorly generalize in scenarios different from the training one (the so-called "domain shift" issue [41]).

In this paper, we are interested in learning representations that are discriminative for the supervised learning task of interest, while being invariant to certain specified *biased attributes* of the data. By "biased attribute", we mean an inherent bias of the dataset, which is assumed to be known and follows a certain distribution during training. At test time, the distribution of such attribute may abruptly change, thus tampering the generalization capability of the model and affecting its performance for the given task [4, 33, 21].

One intuitive example is provided in Figure 1: we seek to

train a *shape classifier*, but each shape has a distinct color – the biased attribute. Unfortunately, a model can fit the training distribution by discriminating either the color or the shape. Among the two options, we are interested in the latter only, because the first one does not allow generalizing to shapes with different colors. Thus, if we were capable of learning a classifier while unlearning the color, we posit that it would better generalize to shapes with arbitrary colors. Like other prior works [29, 33, 21, 4], we operate in a scenario where the labels of biased attributes are assumed to be known. An example of application domain in which the hypothesis of having known labels for the bias holds, is algorithmic fairness [24, 14, 46, 44], where the user specifies which attributes the algorithm has to be invariant to (*e.g.*, learning a face recognition system which is not affected by gender or ethnicity biases).

In this paper, we tackle this problem through the lens of information theory. Since mutual information can be used to quantify the nonlinear dependency of the learned feature space with respect to the dataset bias, we argue that a good strategy to face the aforementioned problem is minimizing the mutual information between the learned representation and the biased attributes. This would result in a data representation that is statistically independent from the specified bias, and that, in turn, would generalize better.

Unfortunately, the estimation of the mutual information is not a trivial problem [35]. In the context of representation learning, two bodies of work proposed solutions to the problem of learning unbiased representations via information theoretic measures: one that relies on adversarial training [4, 21], and one based on variational inference [33]. Adversarial methods [4, 21] learn unbiased representations by "fooling" a classifier trained to predict the attribute from the learned representation. Such condition is argued to be a proxy for the minimization of the mutual information [21]. Other approaches rely on variational inference, which properly formalizes the prior and the conditional dependences among variables. However, when implementing those methods in practice, approximations need to be done to replace the computationally intractable posterior with an auxiliary distribution, but at the cost of several assumptions of independence among the variables. Moreover, such methods are more problematic to scale to complex computer vision tasks, and have been applied mostly on synthetic or toy datasets [29, 33].

Due to the aforementioned difficulties, in this paper, we leverage the mathematical soundness of mutual information to design a computational pipeline that is alternative to standard adversarial training. We rely on a neural estimator for the mutual information (MINE [6]). This module provides a more reliable estimate of the mutual information [35], while still being fully differentiable and, therefore, trainable via backpropagation [37].

Endowed with this model, we propose a training scheme where we alternate between (i) optimizing the estimator and (ii) learning a representation that is both discriminative for the desired task and statistically independent from the specified bias. A key strength of the proposed approach is that the module that estimates the mutual information is not competing with the feature extractor (differently from existing adversarial methods [21]). For this reason, MINE can be trained until convergence at every training step, avoiding the need to carefully balance between steps (i) and (ii), and guaranteeing an updated estimate of the mutual information throughout the training process. In adversarial methods such as [21], where the estimate for the mutual information is modeled via a discriminator that the feature extractor seeks to fool [15, 16], one cannot train an optimal discriminator at every training iteration. Indeed, if one trains an optimal bias discriminator, the feature extractor will no longer be able to fool it, due to the fact that gradients will become too small [5] – and the adversarial game will not reach optimality. This difference is a key novelty of the proposed computational pipeline, which scores favorably with respect to prior work on different computer vision benchmarks, from color-biased classification to age-invariant recognition of people attributes.

Furthermore, a critical aspect of this line of work [4, 21] is how to balance between learning the desired task and "unlearning" the dataset bias, which is a core, open issue [46]. The training strategy proposed in this paper allows for a very simple solution to this important problem. Indeed, as we will show later in the experimental analysis, a very effective approach is selecting the models whose learned representation distribution has the lowest mutual information with that of the biased attribute. We empirically show that these models are also the ones that better generalize to unbiased settings. Most notably, this also provides us with a simple cross-validation strategy for the hyper-parameters: without using any *validation data*, we can select the optimal model as the one that achieves the best fitting to the data, while better minimizing the mutual information. The importance of this contribution is that, when dealing with biased datasets, also the validation set will likely suffer from the same bias, making hyper-parameter selection a thorny problem. Our proposed method properly responds to this problem, whereas former works have not addressed the issue [21].

2. Related Work

The problem of learning unbiased representations has been explored in several sub-fields. In the following section, we cover the most related literature, with particular focus on works that approach our same problem formulation, highlighting similarities and differences.

In **domain adaptation** [20, 8, 38], the goal is learning representations that generalize well to a (target) do-

main of interest, for which only unlabeled – or partially labeled – samples are available at training time, leveraging annotations from a different (source) distribution. In domain generalization, the goal is to better generalize to unseen domains, by relying on one or more source distributions [34, 28]. Adversarial approaches for domain adaptation [15, 16, 42, 43] and domain generalization [39, 47] are very related to our work: their goal is indeed learning representations that do not contain the domain bias, and therefore better generalize in out-of-distribution settings. Differently, in our problem formulation we aim at learning representations that are invariant towards specific attributes that are given at training time.

A similar formulation is related to the so-called "**algorithmic fairness**" [24]. The problem here is learning representations that do not rely on sensitive attributes (such as, *e.g.*, gender, age or ethnicity), in order to prevent from learning discriminant capabilities towards protected categories. Our methods can be applied in this setting, in order to minimize the mutual information between the learned representation and the sensitive attribute (whose distribution might be biased for what concerns the training set). In these settings, it is important to notice that a "fairer" representation does not necessarily generalize better than a standard one: the trade-off between accuracy and fairness is termed "fairness price" [24, 14, 46, 44].

There is a number of works that share our same goal and problem formulation. Alvi et al. [4] learn unbiased representations through the minimization of a confusion loss, learning a representation that does not inherit information related to specified attributes. Kim et al. [21] propose to minimize the mutual information between learned features and the bias. However, they face the optimization problem through standard adversarial training: in practice, in their implementation [2], the authors rely on a discriminator trained to detect the bias as an estimator for the mutual information, and learn unbiased representations by trying to fool this module, drawing inspiration from the solution proposed by Ganin and Lempitsky [15] for domain adaptation. Moyer et al. [33] also introduce a penalty term based on mutual information, to achieve representations that are invariant to some factors. In contrast with related works [4, 21, 33], it shows that adversarial training is not necessary to minimize such objective, and the problem is approached in terms of variational inference, relying on Variational Auto-Encoders (VAEs [23]). Closely related to Moyer et al., other works [29, 12] impose a prior on the representation and the underlying data generative factors (*e.g.*, feature vectors are distributed as a factorized Gaussian).

Our proposed solution provides several advantages to existing adversarial approaches [4, 21] and VAE based ones [33]. With respect to adversarial strategies, our method has the advantage of relying on a module estimating the mutual information [6] that is not competing with the network trained to learn an unbiased representation. In our computational pipeline, we do not learn unbiased representation by "*fooling*" the estimator, but by minimizing the information that it measures. The difference is subtle (for instance, we also end up approaching a minmax optimization problem), but brings a crucial advantage: in standard adversarial methods, the discriminator (estimator) cannot be trained until convergence at every training step, otherwise gradients flowing through it would be close to zero almost everywhere in the parameter space [5], preventing from learning an unbiased representation. In our case, the estimator can be trained until convergence at every training step, improving the quality of its measure without any drawbacks. Furthermore, our solution can easily scale to large architectures (*e.g.*, for complex computer vision tasks) in a straightforward fashion. While this is true also for adversarial methods [4, 21], we posit that it might not be the case for methods based on VAEs [33], where one has to simultaneously train a feature extractor/encoder and a decoder.

3. Problem Formulation

We operate in a setting where data are shaped as triplets $(\mathbf{x}, \mathbf{y}, \mathbf{c})$, where $\mathbf{x}$ represents a generic datapoint, $\mathbf{y}$ denotes the ground truth label related to a task of interest and $\mathbf{c}$ encodes a vector of given attributes. We are interested in learning a representation $\mathbf{z}$ of $\mathbf{x}$ that allows performing well on the given task, with the constraint of not retaining information related to $\mathbf{c}$. In other words, we desire to learn a model that, when fed with $\mathbf{x}$, produces a representation $\mathbf{z}$ which is maximally discriminative with respect to $\mathbf{y}$, while being invariant with respect to $\mathbf{c}$.

In this work, we formalize the invariance of $\mathbf{z}$ with respect to $\mathbf{c}$ through the lens of information theory, imposing a null mutual information I. Specifically, we constrain the discriminative training (finalized to learn the task of interest) by imposing $I(Z, C) = 0$, where Z and C are the random variables associated with $\mathbf{z}$ and $\mathbf{c}$, respectively. In formulæ, we obtain the following constrained optimization

$$\min_{\theta,\psi} \mathcal{L}_{task}(\theta, \psi), \quad s.t. \quad I(Z, C) = 0 \tag{1}$$

where θ and ψ define the two sets of parameters of the objective $\mathcal{L}_{task}$, which can be tailored to learn the task of interest. With θ, we refer to the trainable parameters of a module g_θ that maps a datapoint $\mathbf{x}$ into the corresponding feature representation $\mathbf{z}$ (that is, $\mathbf{z} = g_\theta(\mathbf{x})$). With ψ, we denote the trainable parameters of a classifier that predicts $\tilde{\mathbf{y}}$ from a feature vector $\mathbf{z}$ (that is, $\tilde{\mathbf{y}} = f_\psi(\mathbf{z})$). The constraint $I(Z, C) = 0$ does not depend upon ψ, but only upon θ, since $\mathbf{z}$ obeys to p_Z and $\mathbf{z} = g_\theta(\mathbf{x})$.

In order to optimize the objective in (1), we must adopt an estimator of the mutual information. Before detailing

our approach, in the following paragraph we cover the background required for a basic understanding of mutual information estimation, with focus on the path we pursue in this work.

Background on information theory. The mutual information between two random variables X, Z is given by

$$I(X,Z) = \int p_{X,Z}(x,z) \log \frac{p_{X,Z}(x,z)}{p_X(x) \cdot p_Z(z)} dxdz,$$

where $p_{X,Z}$ denotes the joint probability of the two variables and p_X, p_Z represent the two marginals. As an alternative to covariance and other linear indicators of statistical dependence, mutual information can account for generic inter-relationships between X, Z, going beyond simple correlation [10, 9].

The main drawback with mutual information relates to its difficult computation, since the probability distributions p_X, p_Z and $p_{X,Z}$ are not known in practice. Recently, a general purpose and efficient estimator for mutual information has been proposed by Belghazi et al. [6]. They propose a neural network based approximation to compute the following lower bound for the mutual information I:

$$\begin{aligned} \mathcal{L}_{ne} \coloneqq \mathbb{E}_{(\mathbf{z},\mathbf{c})\sim\hat{p}_{Z,C}}[T_\phi(\mathbf{z}|\mathbf{c})] + \\ - \log \mathbb{E}_{\mathbf{z}\sim\hat{p}_Z, \tilde{\mathbf{c}}\sim\hat{p}_C}[\exp T_\phi(\mathbf{z}|\tilde{\mathbf{c}})] \end{aligned} \tag{2}$$

When implementing T_ϕ as a feed-forward neural network, the maximization in Eq. (2) can be efficiently solved via backpropagation [6].

As a result, we can approximate $I(X,Z)$ with $\widehat{I}_\phi(X,Z)$, the so-called "Mutual Information Neural Estimator" (MINE [6]). An appealing aspect of MINE is its fully differentiable nature, that enables end-to-end optimization of objectives that rely on mutual information computations.

4. Method

In the following, we detail how we approach Eq. (1), both in terms of theoretical foundations and practical implementation.

4.1. Optimization problem

In order to proceed with a more tractable problem, we consider the Lagrangian of Eq. (1)

$$\min_{\theta,\psi} \mathcal{L} \coloneqq \mathcal{L}_{task}(\theta,\psi) + \lambda I(Z,C) \tag{3}$$

where the first term is a loss associated with the task of interest, whose minimization ensures that the learned representation is sufficient for our purposes. The second term is the mutual information between the learned representation and the given attributes. The hyper-parameter λ balances the trade-off between optimizing for a given task and minimizing the mutual information.

Concerning the first term of the objective, we will consider classification tasks throughout this work, and thus we assume that our aim is minimizing the cross-entropy loss between the output of the model $\tilde{\mathbf{y}}$ and the ground truth $\mathbf{y}$.

$$\mathcal{L}_{task} \coloneqq \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i^T \log s(\tilde{\mathbf{y}}_i) \tag{4}$$

where s is the softmax function and N is the number of given datapoints.

Concerning the second term of the objective in Eq. (1), as already mentioned, the analytical formulation of the mutual information is of scarce utility to evaluate $I(Z,C)$. Indeed, we do not explicitly know the probability distributions that the learned representation and the attributes obey to. Therefore, we need an estimator for the mutual information $\widehat{I}(Z,C)$, with the requirement of being differentiable with respect to the model parameters θ.

In order to attain our targeted goal, we take advantage of the work by Belghazi et al. [6], and exploit a second neural network T_ϕ ("statistics network") to estimate the mutual information. We therefore introduce the additional loss function $\mathcal{L}_{ne}$ (Eq. (2)) that, once maximized, provides an estimate of the mutual information

$$\widehat{I}_{ne}(Z,C) = \max_\phi \mathcal{L}_{ne}. \tag{5}$$

In Eq. (2), the notation $\hat{p}$ reflects that we rely on the empirical distributions of features and attributes, the operator "|" indicates vector concatenations and "ne" stands for "neural estimator" [6]. The loss $\mathcal{L}_{ne}$ also depends on θ, since Eq. (2) depends on $\mathbf{z}$. Combining the pieces together, we obtain the following problem:

$$\begin{aligned} \min_{\theta,\psi}\{\mathcal{L}_{task}(\theta,\psi) + \lambda \widehat{I}_{ne}(Z,C)\} = \\ = \underbrace{\min_{\theta,\psi}\{\mathcal{L}_{task}(\theta,\psi) + \lambda \underbrace{\max_\phi \mathcal{L}_{ne}(\phi,\theta)}_{\text{MI estimation}}\}}_{\text{Representation learning}} \end{aligned} \tag{6}$$

Intuitively, the inner maximization problem ensures a reliable estimate of the mutual information between the learned representation and the attributes. The outer minimization problem is aimed at learning a representation that is at the same time optimal for the given task and unbiased with respect to the attributes.

4.2. Implementation Details

Concerning the modules introduced in Section 3, we implement the feature extractor g_θ (which computes features

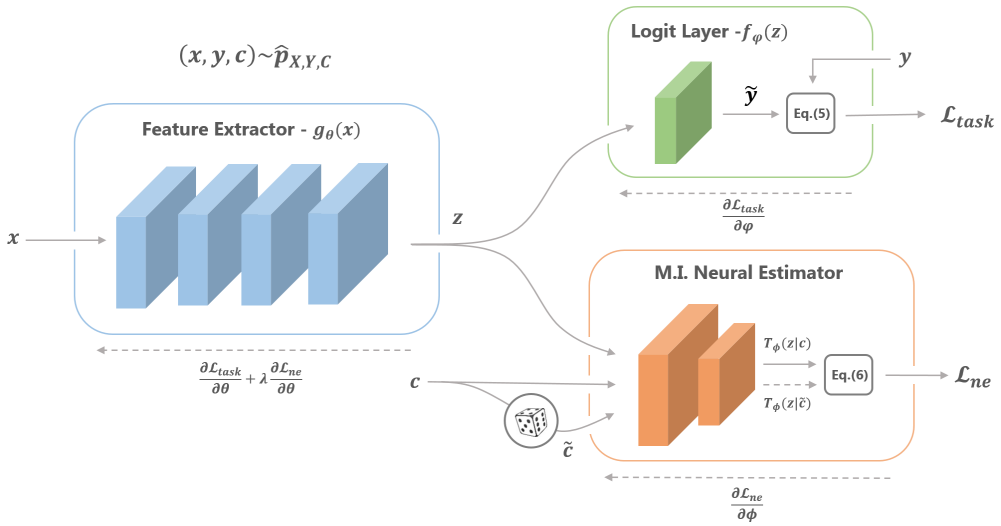

Figure 2: **Model overview.** The neural network devised for the given task is the concatenation of the blue module (feature extractor g_θ) and the green module (logit layer f_ψ). Solid lines indicate the forward flow, dashed lines indicate gradient backpropagations. The feature extractor takes in input samples $\mathbf{x}$ and outputs feature vectors $\mathbf{z}$. The logit layer takes in input the feature vectors and outputs predictions $\tilde{\mathbf{y}}$. To optimize for the given task, these modules can be trained by minimizing the cross-entropy between predictions and labels $\mathbf{y}$. The orange module [6] estimates the mutual information between the feature vectors $\mathbf{z}$ and the attributes $\mathbf{c}$. To estimate the mutual information, T_ϕ processes the concatenation of feature vectors and attributes from the joint distribution and the marginals. Following Belghazi et al. [6], we approximate sampling from the marginal by shuffling the batch of attributes ($\tilde{\mathbf{c}}$). The estimation of the mutual information is the maximum w.r.t. ϕ of the output of the orange module $\mathcal{L}_{ne}$.

$\mathbf{z}$ from datapoints $\mathbf{x}$) and the classifier f_ψ (which predicts labels $\tilde{\mathbf{y}}$ from $\mathbf{z}$) as feed-forward neural networks. The classifier f_ψ is implemented as a shallow logit layer to accomplish predictions on the task of interest. As already mentioned, the model T_ϕ is also a neural network; it accepts in input the concatenation of feature vectors $\mathbf{z}$ and attribute vectors $\mathbf{c}$, and through Eq. (2) allows estimating the mutual information between the two random variables. The nature of the modules allow to optimize the objective functions in (6) via backpropagation [37]. Figure 2 portrays the connections between the different elements, and how the losses (4) and (2) originate.

A crucial point that needs to be addressed when jointly optimizing the two terms of Eq. (6) is that, while the distribution of the attributes $\widehat{p}_C$ is static, the distribution of the feature embeddings $\widehat{p}_Z$ depends on θ, which changes throughout the learning process. For this reason, the mutual information estimator needs to be constantly updated during training, because an estimate $\widehat{I}_{ne}(Z_t, C)$, associated with θ_t at step t, is no longer reliable at step $t+1$. To cope with this issue, we devise an iterative procedure where, prior to every gradient descent update on (θ, ψ), we update MINE on the current model, through the inner maximizer in Eq. (6). This guarantees a reliable mutual information estimation. One key difference with standard adversarial methods[21, 4] is that we can train MINE until convergence prior to each gradient descent step on the feature extractor, without the risk of obtaining gradients whose magnitude is close to zero [5], since our estimator is not a discriminator

Algorithm 1 Learning Unbiased Representations

Input: Dataset $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}, \mathbf{c}^{(i)})\}_{i=1}^N$, initialized weights θ_0, ψ_0, ϕ_0, learning rates α, η, hyper-parameters λ, K, T.
Output: learned weights θ, ψ
Initialize: $\theta \leftarrow \theta_0$, $\psi \leftarrow \psi_0$, $\phi \leftarrow \phi_0$
for t $= 1, ..., T$ **do**
  for k $= 1, ..., K$ **do** (estimate MI)
    sample mini-batches $\{(\mathbf{x}^{(i)}, \mathbf{c}^{(i)})\}_{i=1}^m$, $\{\tilde{\mathbf{c}}^{(i)}\}_{i=1}^m$
    evaluate $\mathcal{L}_{ne}$ (Eq. (2))
    $\phi \leftarrow \phi + \eta \nabla_\phi \mathcal{L}_{ne}$
  sample mini-batches $\{(\mathbf{x}^{(i)}, \mathbf{x}^{(i)}, \mathbf{c}^{(i)})\}_{i=1}^n$, $\{\tilde{\mathbf{c}}^{(i)}\}_{i=1}^n$
  evaluate $\mathcal{L}_{task}$ (Eq. (4)) and $\mathcal{L}_{ne}$ (Eq. (2))
  $\theta \leftarrow \theta - \alpha \nabla_\theta (\mathcal{L}_{task} + \lambda \mathcal{L}_{ne})$
  $\psi \leftarrow \psi - \alpha \nabla_\psi \mathcal{L}_{task}$

(being the mutual information unbounded, sometimes gradient clipping is actually beneficial [6]). The full training procedure is detailed in Algorithm 1.

Training techniques. We list some techniques that we could appreciate to generally increase the stability of the proposed training procedure. While code and hyper-parameters can be found in the Supplementary Material, we believe that the reader can benefit from the discussion.

(a) Despite MINE [6] can estimate the mutual information between continuous random variables, we observed that the estimation is eased (in terms of speed and stability)

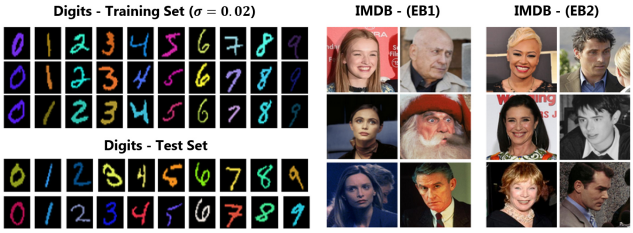

Figure 3: *Left:* digit examples for each class from training (here with $\sigma = 0.02$) and test set. *Right:* Women and Men images from the two splits of the training set of the IMDB dataset.

if the attribute labels $\mathbf{c}$ are discrete. ***(b)*** We observed an increased stability in training MINE [6] for lower-dimensional representations $\mathbf{z}$ and attributes $\mathbf{c}$. For this reason, as we will discuss in Section 5, feature extractors with low-dimensional embedding layer are favored. ***(c)*** The feature extractor g receives gradients related to both $\mathcal{L}_{task}$ and $\mathcal{L}_{ne}$: since the mutual information is unbounded, the latter may dominate the former. Following Belghazi et al. [6], we overcome this issue via gradient clipping (we refer to original work for details). ***(d)*** We observed that training MINE requires large mini-batches: when this was unfeasible due to memory issues, we relied on gradient accumulation. ***(e)*** We observed that using vanilla gradient descent over Adam optimizer [22] eases training MINE [6] in most of our experiments.

5. Experiments

In the following, we show the effectiveness of models trained via Algorithm 1 in a series of benchmarks. First, we report results related to the setup proposed by Kim et al. [21] – learning to recognize color-biased digits without relying on color information. Next, we show that our proposed solution can scale to higher-capacity models and more difficult tasks, through the IMBD benchmark [4, 21], where the goal is classifying people gender without relying on the age bias. Finally, we show that our method can also be applied as it is to learn "fair" classifiers, by training models on the two standard benchmarks for algorithmic fairness [1, 3].

5.1. Digit Recognition

Experimental setup. Following the setting defined by Kim et al. [21], we consider a digit classification task where each digit, originally from MNIST [27], shows an artificially induced color bias. More specifically, in the training set (with $60,000$ samples), digit colors are drawn from Gaussian distributions, whose mean values are different for each class. In the test set (with $10,000$ samples), digits show random colors. The benchmark is designed with seven different standard deviation values σ (equally spaced between 0.02 and 0.05): the lower the value, the more difficult the task, since the model can fit the training set by recognizing colors instead of shapes, thus poorly generalizing (see Figure 3). To extract the color information (the attribute $\mathbf{c}$, recalling notation from Section 3), the maximum pixel value is encoded in a binary vector with 24-bit (8 bits per channel). Since the background is always black, the maximum value reflects the digit color.

Concerning the model, we exploit a convolutional neural network [26] with architecture *conv-pool-conv-pool-fc-fc-softmax*. The output of the second fully connected layer ($\mathbf{z}$) is given in input to both the logit layer and MINE (Figure 2). The architecture of the statistics network T_ϕ in MINE is a multi-layer perceptron (MLP) with 3 layers. More architectural details can be found in the Supplementary Material. We compare models trained via Algorithm 1 with the solutions proposed by Kim et al. [21] and Alvi et al. [4], averaging across 3 runs and using accuracy as a metric. Before comparing against related work, we discuss how crucial hyper-parameters can be selected in our setting.

Hyper-parameter choice. We discuss in the following the model behavior as we modify λ, that governs the trade-off between learning a task and minimizing the mutual information between features and attributes.

Figure 4 reports the evolution of mutual information estimation (left), accuracy on test samples (middle) and accuracy on training sample (right) for models trained with $\lambda = 0.0, 0.5, 1.0$ in blue, orange and green, respectively, for $\sigma = 0.03, 0.045$ (top and bottom, respectively). It can be observed that the mutual information between embeddings $\mathbf{z}$ and color attributes $\mathbf{c}$ can be reduced by increasing λ. Importantly, this results in a significantly higher accuracy on (unbiased) test samples. The importance of this result is twofold: on the one hand, it is a proof of concept of the intuition that lowering the mutual information does help generalizing to unbiased sources; on the other, it provides us a cross-validation strategy to pick a proper λ value (the one that allows minimizing the mutual information more efficiently). As can be observed in the plots on the right, the training procedure becomes more unstable when we increase λ. Therefore, in order to select the proper hyper-parameter, we can choose the highest λ value that allows the model fitting the data (*i.e.*, minimizing $\mathcal{L}_{task}$) and reducing the mutual information (*i.e.*, minimizing $\mathcal{L}_{ne}$).

Another important hyper-parameter is the number of iterations used to train MINE [6] prior to each gradient update on the feature extractor (K in Algorithm 1). We observed that, the higher the number of iterations the better (we set $K = 80$). This was expected, because the MI estimation becomes more reliable and therefore the removal of bias information is more accurate. The reader can refer to Figure 3 in the Supplementary Material for quantitative results.

Comparison with related work. We report in Table 1 the comparison between our method with $\lambda = 1.0$ and related works [21, 4]. We can observe consistently improved re-

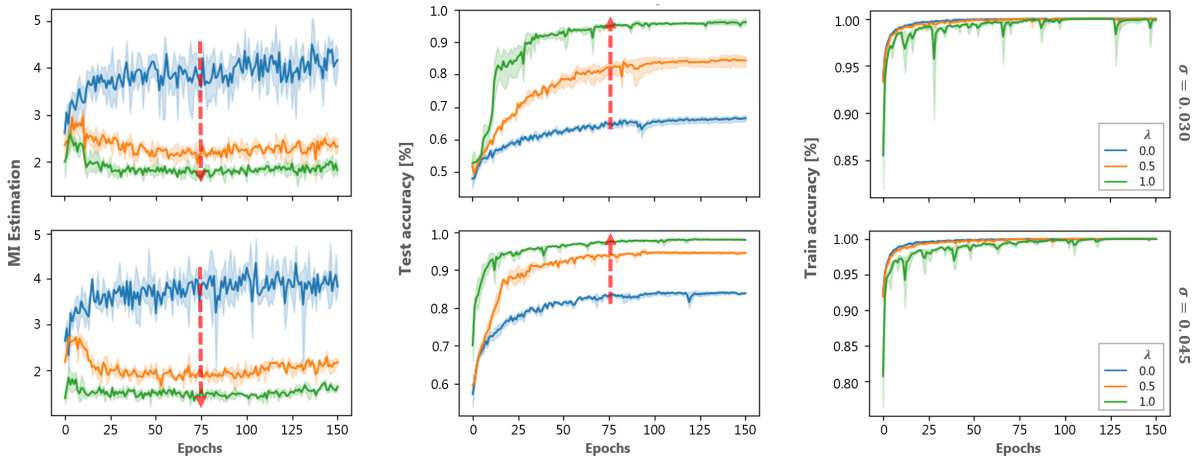

Figure 4: **Digit experiment – ablation study.** Evolution of mutual information estimation (left), test accuracy (middle) and training accuracy (right) for models trained on digits with $\sigma = 0.03$ and $\sigma = 0.045$ (top and bottom, respectively). Models are trained with Algorithm 1 with $\lambda = 0.0$ (baseline, blue), $\lambda = 0.5$ (orange) and $\lambda = 1.0$ (green). Increasing the value of the hyper-parameter λ allows reducing the mutual information between the learned representation (Z) and the attributes (C). In turn, models better generalize to unbiased samples (test set). Further plots in the Supplementary.

Digit experiment

	Color variance						
Training	$\sigma = 0.020$	$\sigma = 0.025$	$\sigma = 0.030$	$\sigma = 0.035$	$\sigma = 0.040$	$\sigma = 0.045$	$\sigma = 0.050$
ERM ($\lambda = 0.0$)	0.476 ± 0.005	0.542 ± 0.004	0.664 ± 0.007	0.720 ± 0.010	0.785 ± 0.003	0.838 ± 0.002	0.870 ± 0.001
Alvi et al. [4]	0.676	0.713	0.794	0.825	0.868	0.890	0.917
Moyer et al. [33]	0.717	0.864	0.883	0.885	0.887	0.893	0.914
Kim et al. [21]	0.818	0.882	0.911	0.929	0.936	0.954	0.955
Ours ($\lambda = 1.0$)	0.864 ± 0.052	0.925 ± 0.020	0.959 ± 0.008	0.973 ± 0.003	0.975 ± 0.001	0.980 ± 0.001	0.982 ± 0.001

Table 1: **Digit experiment – comparison with related work.** Experimental results on colored digit classification for different levels of variance (σ) in the color distribution. The first row reports results related to models trained via standard Empirical Risk Minimization (ERM). Below we report results obtained by competitor methods [4, 33, 21]. The last row reports results achieved with our method (with $\lambda = 1.0$.)

sults in all splits (different σ's). We emphasize that our method is more effective as the bias is more severe (small σ's). It is also important to stress that other works [21, 33, 4] do not introduce any strategy to tune the hyper-parameters, whereas in this work the hyper-parameter search is efficiently resolved. Furthermore, the authors do not report any statistics around their results (*e.g.*, average and standard deviation across different runs), making a fair comparison difficult.

IMDB experiment

	Train on EB1		**Train on EB2**	
Method	**EB2**	**Test**	**EB1**	**Test**
ERM ($\lambda = 0.0$)	0.650 ± 0.020	0.849 ± 0.007	0.576 ± 0.013	0.708 ± 0.008
Alvi et al. [4]	0.637 [21]	0.856 [21]	0.573 [21]	0.699 [21]
Kim et al. [21]	0.680	0.867	0.642	0.745
Ours ($\lambda = 0.5$)	0.691 ± 0.010	0.876 ± 0.010	0.651 ± 0.036	0.762 ± 0.022

Table 2: Experimental results from IMDB gender classification problem. The first row reports results obtained by setting $\lambda = 0.0$ (ERM baseline). The last row reports results obtained with our method (*Ours*); Each column reports results associated with the indicated test set.

5.2. IMDB: Removing the Age Bias

Experimental setup. Following related works [4, 21], we consider the IMDB dataset [36] as benchmark. It contains cropped images of celebrity faces with ground truth annotations related to gender and age. Alvi et al. [4] consider two subsets of the training set that are severely biased for what concerns age: the EB1 ("Extreme Bias") split ($36,003$ samples) only contains images of women with an age in the range 0-30, and men who are older than 40; vice versa, the EB2 split ($16,799$ samples) only contains images of men with an age in the range 0-30, and women older than 40 (see Figure 3). The test set ($22,468$ samples) contains faces without any restrictions on age/gender (uniformly sampled). The goal here is learning an age-agnostic model, to overcome the bias present in the dataset.

Following previous work [4, 21], we encode the age attribute (our biased attribute, $\mathbf{c}$) using bins of 5 years, via one-hot encoding. We use a ResNet-50 [17] model pre-trained on ImageNet [13] as classifier, modified with a 128-dimensional fully connected layer before the logit layer. This narrower embedding serves as our $\mathbf{z}$, and the

Fairness experiment				
	Adult dataset		German dataset	
Method	Acc ↑	EO ↓	Acc ↑	EO ↓
FERM [14]	0.81	0.01	0.73 ± 0.04	0.05 ± 0.03
NN [31]	0.84	0.14	0.74 ± 0.04	0.47 ± 0.19
NN + χ [31]	0.83	0.03	0.73 ± 0.03	0.25 ± 0.14
LAFTR [30]	0.84	0.10	–	–
Ours ($\lambda = 0.5$)	0.85 ± 0.01	0.03 ± 0.02	0.74 ± 0.04	0.06 ± 0.05

Table 3: **Fairness experiments** – We compare against results on the two datasets as reported in [14, 31, 30]. For accuracy, the higher the better. For EO, the lower the better (*i.e.*, the "fairer"). Our results were averaged across 10 different runs.

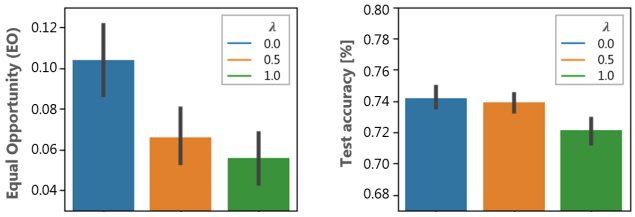

Figure 5: The two considered metrics vary as we modify the hyper-parameter λ on the German dataset. EO (*Left*) is significantly reduced as we set higher values of λ. Vice versa, test accuracy (*Right*) is only slightly affected.

reduced dimension eases the estimation of the mutual information, while not causing any detrimental effect in terms of accuracy. For each split (EB1 and EB2), we train the model through Algorithm 1 and evaluate it on the test set and on the split not used for training. We followed the same procedure detailed in Section 5.1 to choose the hyper-parameter λ; we set $K = 40$. We compare our results with the ones published by related works [4, 21], using accuracy as a metric. We limited the training sets to only $2,000$ samples: this choice was due to the fact that with the whole training sets we could observe baselines ($\lambda = 0.0$) significantly higher than published results [21], whereas they are comparable for models trained on a subset.

Results. Table 2 reports our results. In all our experiments, we observe accuracy improvements with respect to the baseline ($\lambda = 0.0$). In general, training on one split and testing on the other is more challenging than testing on the (neutral) test set, as confirmed by the baseline results (ERM, first row). In all the different protocols, our method (last row) has superior performance than Alvi et al. [4], and comparable or superior performance with Kim et al. [21].

These results confirm that our method can effectively remove biased, detrimental information even when modeling more complex data with higher-capacity models. In this case though, the improvements are more limited than the ones we showed in the digit experiment. One of the reasons might be that age and gender cannot be decoupled as efficiently as shape and color. In other words, removing age information may not necessarily increase accuracy.

5.3. Learning Fair Representations

Experimental setup. We explored the potentiality of our method in the context of algorithmic fairness with the popular UCI datasets Adult [1] and German [3]. Both datasets contains tabular data with categorical and continuous attributes: Adult has $\sim 48,000$ US adult Census data samples and the goal is to predict whether the person has a annual salary $> 50K\$$; German is composed of $1,000$ samples of bank customer descriptions and the binary, ground truth label is the risk degree associated with a customer, either good or bad. The goal is to learn a model to solve tasks with the constraint of removing sensitive information about gender in Adult and customer age in German. This problem is different with respect to the previous ones: here the invariance towards sensitive attribute does not imply a better generalization on the test set as it happens with, *e.g.*, digit recognition. The removal of the protected attribute is done for the sake of learning a fair representation [24, 14, 46, 44].

Following previous works [30, 14], we implemented the feature extractor as single-layer MLP's. Additional details can be found in the Supplementary Material. We evaluate accuracy and "equal opportunity" (EO)[1] as comparison metrics, averaging across 10 different runs. The goal is to find a balance between reducing EO (*i.e.*, learning a fairer representation) without observing a too severe decrease in accuracy.

Results. In Figure 5, we show how the performance varies when increasing λ from 0 (standard Empirical Risk Minimization) to 1 for models trained on German. It can be observed that our method allows training fairer models (*i.e.*, reduced EO), while maintaining a good performance on test. For $\lambda = 0.5$, the fairness price is close to zero (*i.e.*, the accuracy does not decrease), while the fairness is substantially improved. We report the comparison with related works in Table 3, for both datasets. These results show that our method can be effectively used to tackle algorithmic fairness: we achieve a favorable fairness trade-off, matching or exceeding test accuracy while keeping EO lower than the competitor methods.

6. Conclusions

We propose a training procedure to learn representations that are not biased towards dataset-specific attributes in an alternative fashion to existing adversarial approaches. We leverage a neural estimator for the mutual information [6], devising a method that can be easily implemented in arbitrary architectures and provides a robust strategy for hyper-parameter tuning. We show competitive results on benchmarks ranging from computer vision [4, 21, 33] to fair representation learning [30, 14, 31].

[1] Equal Opportunity measures the discrepancy between the TP rates of "protected" and "non-protected" populations. Here, EO $= |\text{TP}(\text{young}) - \text{TP}(\text{not young})|$.

References

[1] Adult census data set. https://archive.ics.uci.edu/ml/datasets/adult.

[2] Code for "learning not to learn: Training deep neural networks with biased data". https://github.com/feidfoe/learning-not-to-learn.

[3] Statlog (german credit data) data set. \https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data).

[4] Mohsan S. Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *European Conference on Computer Vision (ECCV) Workshops*, 2018.

[5] Martín Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[6] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[7] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, Aug. 2013.

[8] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 120–128, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[9] Jacopo Cavazza, Pietro Morerio, and Vittorio Murino. Scalable and compact 3d action recognition with approximated rbf kernel machines. *Pattern Recognition*, 93:25 – 35, 2019.

[10] Jacopo Cavazza, Andrea Zunino, Marco San Biagio, and Vittorio Murino. Kernelized covariance for action recognition. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 408–413. IEEE, 2016.

[11] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537, 2011.

[12] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi, and Richard S. Zemel. Flexibly fair representation learning by disentanglement. *CoRR*, abs/1906.02589, 2019.

[13] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. Imagenet: A large-scale hierarchical image database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[14] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

[15] Yaroslav Ganin and Victor S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1180–1189, 2015.

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1), Jan. 2016.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[20] Hal Daumé III and Daniel Marcu. Domain adaptation for statistical classifiers. *CoRR*, abs/1109.6341, 2011.

[21] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[23] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[24] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*. 2012.

[26] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[27] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[28] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. *The IEEE International Conference on Computer Vision (ICCV)*, 2017.

[29] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The Variational Fair Autoencoder. *arXiv e-prints*, page arXiv:1511.00830, Nov 2015.

[30] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. *CoRR*, abs/1802.06309, 2018.

[31] Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, pages 4382–4391, 2019.

[32] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[33] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9084–9093. Curran Associates, Inc., 2018.

[34] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 10–18, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[35] Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. 2019.

[36] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)*, July 2016.

[37] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Learning Internal Representations by Error Propagation, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

[38] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 213–226, Berlin, Heidelberg, 2010. Springer-Verlag.

[39] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*, 2018.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[41] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1521–1528, Washington, DC, USA, 2011. IEEE Computer Society.

[42] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[43] Riccardo Volpi, Pietro Morerio, Silvio Savarese, and Vittorio Murino. Adversarial feature augmentation for unsupervised domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[44] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Adversarial removal of gender from deep image representations. *CoRR*, abs/1811.08489, 2018.

[45] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2018.

[46] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. *CoRR*, abs/1801.07593, 2018.

[47] Andrea Zunino, Jacopo Cavazza, Riccardo Volpi, Pietro Morerio, Andrea Cavallo, Cristina Becchio, and Vittorio Murino. Predicting intentions from motion: The subject-adversarial adaptation approach. *International Journal of Computer Vision*, Sep 2019.